PULSE DETECTION IN HORMONE DATA: SIMPLIFIED, EFFICIENT ALGORITHM

Peter J. Munson and David Rodbard
Laboratory of Theoretical and Physical Biology, National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Md 20892

KEY WORDS: Peak detection, deconvolution, endocrinology, pituitary hormones

INTRODUCTION

An important problem in endocrinology is the determination of the number of peaks or episodes of secretion of a pituitary hormone from a series of measurements of its circulating levels. It is known for some hormones, such as luteinizing hormone (LH), that episodic, pulsatile release can have nearly the opposite effect as a steady, tonic release of the same amount of hormone. Several groups (1-5) have attempted to characterize such data sets by first ascertaining the number of underlying "peaks" or "events" giving rise to the data. The problem is complicated by the mixing and dilution of the secreted hormone in the peripheral blood, the dynamic clearance of the hormone by the kidneys and other organs, and by the often large degree of measurement error inherent in hormone assay measurements.

The earliest workers in the field of endocrinological peak detection used visual, graphical methods to evaluate the presence of a secretory peak. Santen and Bardin proposed to define a peak as a sudden rise from "baseline" of more than 20% (6). Other workers have elaborated these methods, including a robust definition of the baseline (2), Fourier spectrum analysis (5), robust definition of a peak (3), and more recently, numerical deconvolution (3,4). Recently, O'Sullivan and O'Sullivan demonstrated the use of the generalized cross-validation index (GCV) in conjunction with numerical deconvolution for data sets of this sort (7). Diggle and Zeger (8) use maximum likelihood estimation for a modified autoregressive process to characterize such data arising in physiological studies of sheep.

Serious problems remain in this area. First, there is no generally accepted, rigorous, formal definition of what constitutes a secretory event or peak. Thus, it is difficult to move toward an optimal peak detector. Perhaps the best definition that can now be given is that a secretory event is a rare, instantaneous occurrence of variable magnitude, during which a finite amount of hormone is placed in the circulation. For technical reasons, some workers characterize such events as having finite duration, with a gaussian or other shape with respect to time.

A second problem is that the clearance kinetics, i.e. the rate at which a secreted hormone decays in the circulation, is undefined. Here, most workers assume first-order, linear kinetics, i.e. decay according to a single negative exponential-time curve, while admitting the possibility of two or more exponential components. The time constants associated with each component are generally unknown at the outset.

A third problem is the dependence or correlation of the number of observed secretory events with the assumed time constant for exponential decay. That is, many data sets may be equally well represented by many peaks with a short decay time as by few peaks with a long decay time. This dependency is an inherent limitation of the data, regardless of the method of analysis, yet the problem seems to have been largely ignored by workers in this area. We will explore this problem further.

Fourth, the concept of "number of peaks" is problematical. For example, due to the limited time-resolution in the data series there may actually be a large number of very small peaks, closely spaced in time, which are "invisible" to the analysis. Even some of the smaller, biologically irrelevant but visible peaks may affect the peak count. Finally, even when the number of events or peaks can be well-defined and well-estimated, one must take account of the magnitude of the events. A train of uniformly sized, uniformly spaced peaks might have a very different biological meaning than a series consisting of highly variable, unevenly spaced peaks. We discuss this problem and propose an alternative solution to simply estimating the number of peaks.

More work is needed to ascertain the statistical properties of the various peak counting/peak detection methods proposed in the literature. Some workers have characterized the "false positive/false negative" rates for their methods by numerical simulation. Statistically sophisticated methods are often computationally intensive, so that numerical characterization of their properties is infeasible. In order to permit numerical simulation, we chose to consider only computationally feasible methods, sacrificing some degree of optimality. We attempted to incorporate as much realism into the underlying model as possible. We take a "data-analytic" approach, with virtually all parts of our model being suggested by features in illustrative data sets.

AN EXAMPLE

The data set shown in Figure 1A represents the sampled concentration of circulating luteinizing hormone (LH), taken every 5 minutes over a 12 hour interval in a normal male volunteer (9). LH is a primary pituitary hormone controlling gonadal function. We note the appearance of 4 obvious major asymmetric peaks, with apparently exponential tails. These peaks are the primary indication of pulsatility in these data. Included in this data set are several apparent outliers, one possibly resulting from a transposition of two data values.

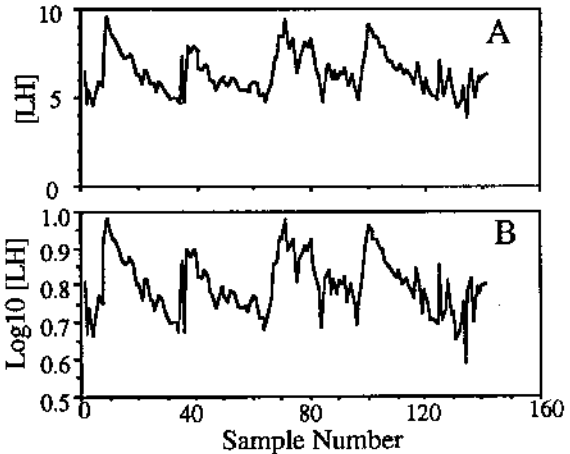

Figure 1. A: LH concentration measured every 5 minutes for a 12 hour period, 144 observations taken from (9). B: Base 10 logarithm of LH concentration.

THE MODEL

The LH data can be characterized as a continuous function with a small number of jumps or step discontinuities, representing the pulses. The apparent exponential tails of each pulse can be linearized by plotting LH on a logarithmic scale, as in Fig. 1B. Ideally, as each pulse is cleared from the circulation by the same mechanism,

Reprinted from the 1989 Proceedings of the Statistical Computing Section of the American Statistical Association

the exponential rate of decay and hence the negative slope should be constant. With this assumption, the data now are representable as a downward sloping straight line broken occasionally by upward jumps, for which we write:

$$y_i = p_1 + \sum_{j=1}^{m_i} p_j - k\,(i-1) + \varepsilon_i;\ i=1, \ldots, n \qquad (1)$$

where y_i is the logarithm of LH concentration at time t_i, p_0 is a baseline or initial value of y at time zero, p_j is the pulse height of the j^{th} pulse in this coordinate system, and ε_i is a normally distributed error term. The sum is taken over the first m_i pulses which occur before the i^{th} time point. Incidentally, the logarithmic transformation, in addition to linearizing the decay component, has the additional advantage of compensating for the commonly observed heteroscedastic measurement error for LH, which has nearly constant coefficient of variation.

We now seek to estimate the number of the upward jumps or "peaks" in the series. Simultaneously, we shall estimate the location and magnitude of each peak, and optionally, the decay rate, k. The relationship between the number of upward jumps and the value of the downward slope, k, is clear. If the slope is too great, too many upward jumps will be required to fit the data. Too gradual a slope will result in too few jumps or "peaks."

One may also think of this model as a "zero-order" spline plus a linear term. A zero-order spline is a series of constant functions separated by points of discontinuity called "knots." Since we are particularly interested in these jump discontinuities, the problem is then to find the appropriate number and location of knots required to represent the data.

PEAK DETECTION ALGORITHM

There are actually three phases to the algorithm we propose to use, corresponding to the estimation of the number, location and magnitude of the peaks. We have deliberately made the model linear in all its parameters, so that the magnitude of the peaks and the decay rate can be estimated in the first phase by linear regression. In the second phase, the location of the peaks will be determined using a variation of stepwise linear regression. In the third phase, determining the number of peaks, we face a problem analogous to deciding when to stop adding terms to a multiple regression model. We shall investigate the use of the generalized cross-validation index (GCV) for this purpose. The use of GCV is partially motivated by the similarity of our model to a spline model, in the context of which the GCV was originally applied(10).

Peak Magnitude

To formulate the problem as a multiple linear regression, consider the set of independent variables x_j, which are to be the basis functions for the spline. that is, we allow for potential pulses at each of the time points. Let x_{ij}, that is x_j at the i^{th} data point, be 0 if i<j and 1 elsewhere. For a particular subset of size m of the n possible peaks, indexed by s, the model may be written

$$y_i = p_1 + \sum_{s=1}^{m} p_{j(s)}\, x_{ij(s)} - k(i-1) + \varepsilon_i;\quad i=1,\ldots,n \qquad (2)$$

We may assume either that the decay rate, k, is known from other data, or that it is to be estimated from the data at hand.

With a known decay rate, we may re-write (2) as

$$y_i^* = y_i + k\,(i-1) = p_1 + \sum_{s=1}^{m} p_{j(s)}\, x_{ij(s)} + \varepsilon_i \qquad (3)$$

The design matrix when this particular set of peaks is included becomes:

$$X_{in} = [\ x_1\ \ x_{j(1)}\ x_{j(2)} \ldots\ x_{j(m)}\ ]$$

where $x_j = (0, 0, \ldots, 0, 1, 1, \ldots, 1)^T$, with the first 1 in the $j+1^{st}$ row of x_j, and the parameter vector $\beta = (p_0, p_{j(1)}, \ldots, p_{j(m)})^T$. When k is to be included as an unknown parameter, we append the vector $(0,1,2,\ldots,n-1)^T$ as the first column of X_{in}.

Peak Location

To determine peak locations, we must determine the particular set of independent variables, x_j, to include in the regression. The variable selection technique for this problem should attempt to explain the greatest residual sum of squares with the least number of parameters, but should also enforce the constraint that the secretory peaks be positive, since a "negative" secretory burst is biologically infeasible. For this reason, we shall investigate only a stepwise addition strategy, guaranteeing satisfaction of the positivity constraint at each step. A more complex strategy including both stepwise deletion and addition would be required to find the best possible subset regression, or the best "positive" subset regression. Such strategies are quite computationally intensive, and not amenable to numerical simulation. In our example, such complex strategies did little to reduce the residual sum of squares, i.e. the simpler strategy recommended here is apparently nearly optimal, in the examples we have tested.

The peaks are "detected" sequentially. To choose the next peak or regressor, we include that variable whose component of the anti-gradient vector is most positive. That is, we use a gradient-directed search for the next peak to include in the regression. This peak is readily determined as the largest component of the gradient, $X'_{out}(y - \hat{y})$, where X_{out} is the matrix of variables not included in the regression, and $\hat{y}$ is the vector of predictions of the current model. Requiring the antigradient component to be positive forces the new peak to be positive, but may still permit other peaks already included to become negative. We check for this latter condition and terminate the search if it becomes true. Alternately, we could have added the peak which explains the largest part of the residual sum-of-squares. This approach can be effectively implemented using the Sweep operator on the sum-of-squares-and-cross-products (SSCP) matrix (10), but it is computationally slightly more expensive than the gradient directed approach, which performed almost as well in terms of the final residual sum of squares.

Number of Peaks

To determine the number of peaks giving rise to the observed data, it is necessary to truncate the stepwise regression procedure at some point. At each step of the regression, the residual sum of squares (SS) will decrease. Likewise, the mean square error calculated as SS/df, will also decrease, approaching an unbiased estimate of the variance of the error term ε. Here, we take the degrees of freedom, df = n-m-2. However, the GCV, defined as SS/df^2, should theoretically decrease to a minimum when the regression model has the best predictive power for the data series, i.e. at the true number of peaks (10). This procedure was also suggested by O'Sullivan (7). Thus, the procedure we adopt is to add terms to the regression until the minimum GCV model is found.

RESULTS

We now illustrate this procedure on the LH data of Figure 1. Figure 2A illustrates the results after a single peak has been found. We include a "peak" (p_1) at the first time point in this and all subsequent models, representing the "baseline" or "initial" value for LH concentration. The gradient vector, plotted as a series, is also shown in Figure 2A. The three largest local maxima indicate potential positions for peaks in subsequent steps. There is obvious inadequacy of the fit of the one-pulse model. Figures 2B-2H show the results after 2 through 5, 10, 15, and 20 peaks

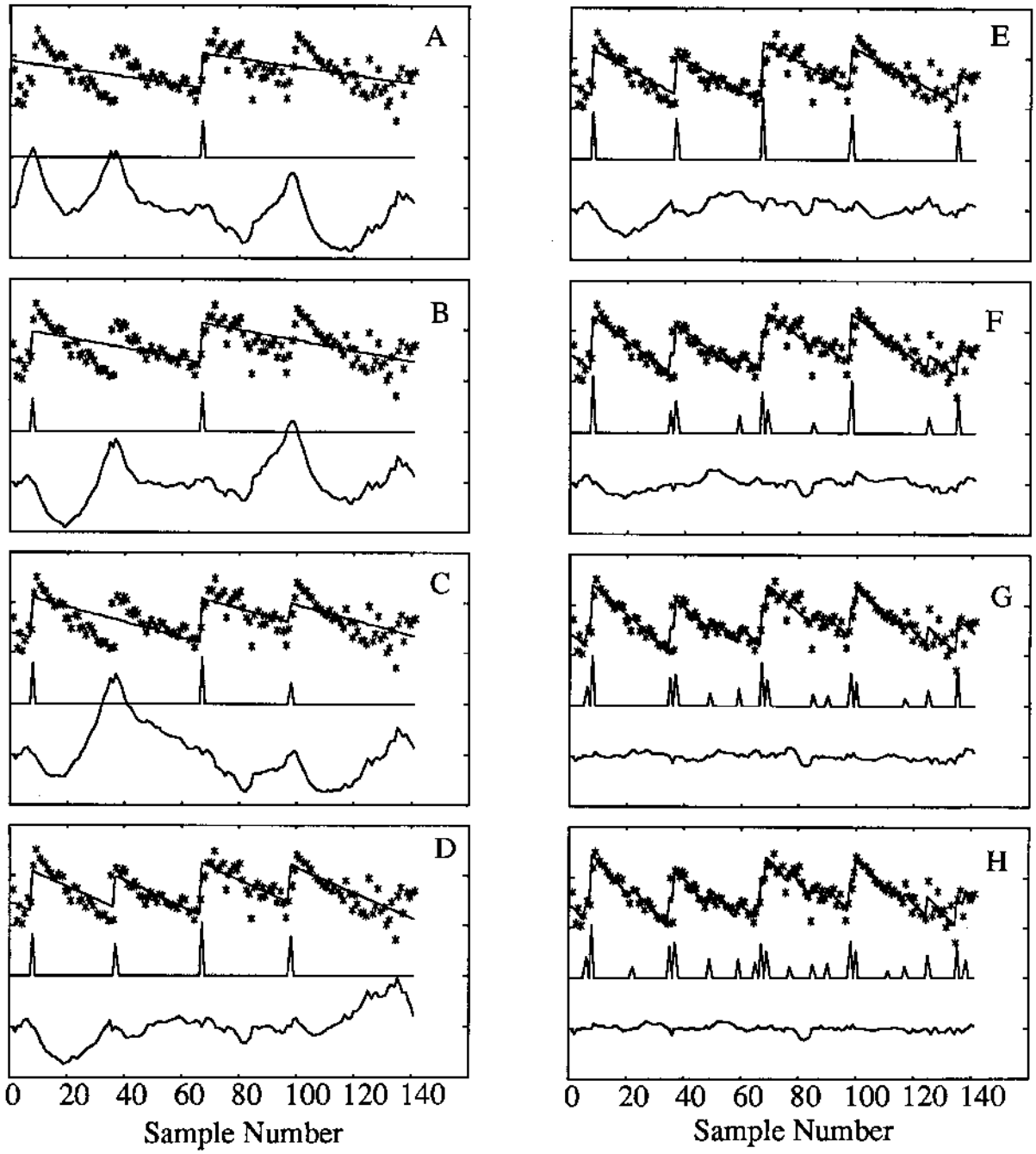

Figure 2A-H. Results of models involving 1,2,3,4,5,10,15,or 20 peaks, respectively. In each panel, Upper trace: Reconstructed data superimposed on actual data (*); Middle trace: Peaks found by current model. Lower trace: Components of gradient vector for each model. Maximum of gradient vector components is chosen as the position of the next peak to add to the model.

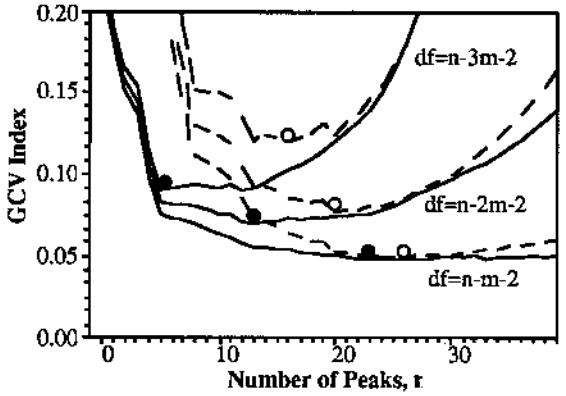

Figure 3. GCV index (=SS/df^2) for models chosen by the new algorithm. Solid lines, Lower trace: df=n-m-2; Middle trace: df=n-2m-2; Upper trace: df=n-3m-2; Dashed lines: decay factor k is fixed at 0.03. (●, ○) indicate positions of minimum of GCV index.

have been detected, respectively. Choice of appropriate model can be made visually or through the use of the GCV index, plotted in Figure 3. Although, there were 4 or 5 visually obvious major peaks in the original series, the GCV minimal model includes 23 peaks. While the 5 peak model did show some evidence of lack of fit and the additional 18 peaks did reduce the SS two-fold, one suspects the physiological reality of those 18 additional peaks. Some of them do not seem very likely to be meaningful upon visual inspection of the data.

Number of Peaks vs. Decay Rate

Figure 4 is a plot of estimated decay rate k vs number of peaks for the LH example, showing a consistent, nearly linear relationship. From Figure 3, we see that there is a very broad range in the acceptable number of peaks, in that the GCV changes negligibly from about m=20 to 40. The corresponding range in value for k is 0.03 to 0.05. Thus, neither estimates of number of peaks nor of the clearance rate are likely to be very precise using this kind of data. One alternative approach would be to estimate the clearance rate from another experiment (e.g., using a bolus injection of hormone), and include the result as a fixed parameter. As an example, we set the value of k=0.03, and recomputed the regression. Now, the GCV minimum occurs at 26 peaks, with a narrower acceptable range from about 22 to 29 (Figure 3, lower trace, dashed lines, open symbol). Thus, external measurements of the decay rate would appear to be userful for estimating the number of peaks.

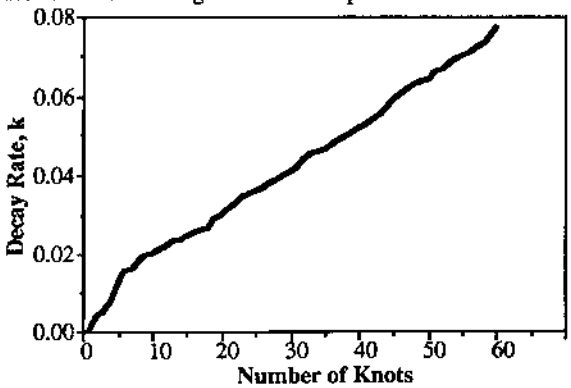

Figure 4. Plot of estimated decay rate, k, vs. m, the number of peaks in the model showing nearly linear relationship.

Numerical Simulation

The results from the previous section are disappointing in two respects. First, there is no clear cut answer to the question "How many peaks?" Second, the best available answer (between 20 and 40) seems much too high when one evaluates Fig. 1 visually. It seems that a reasonable procedure should conclude that there are about 4 large peaks. Many physiologists would suggest, perhaps intuitively, that these four or five sudden major changes in the circulating concentration are the key signal. The remainder of the variability in the signal is probably of less importance. Yet in a model close to the GCV optimum, those four major peaks are barely discernable amongst the other reconstructed peaks (Fig. 2H). To see if this situation was simply the results of random measurement error, we performed a simulation study, using a model with five peaks as the underlying signal. The variance of the error term was set to 0.03 and k was taken at 0.03. Averaged results of 846 runs of the peak-detection algorithm are shown in Figure 5.

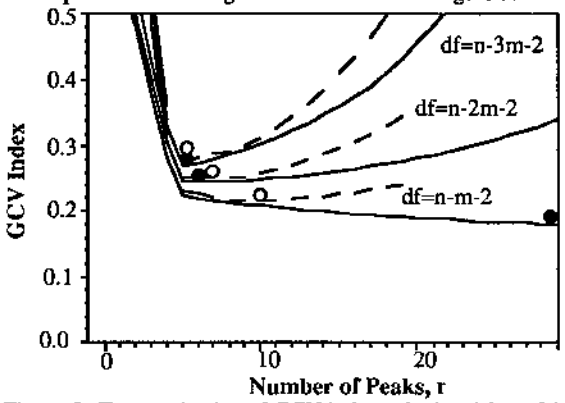

Figure 5. Expected value of GCV index calculated from 846 sets of simulated data using a 5 peak model. Lower, middle and upper traces, symbols as in Figure 4. Solid lines: k is estimated parameter; Dashed lines: k = true value, 0.03.

From these result, we see that the GCV optimal number of peaks is indeed biased (observed value: ≥30, true value: 5). Even when the decay rate, k, is fixed at its true value, the GCV reaches a minimum with 10 peaks (Fig. 5, dashed lower trace, open symbol). When the decay rate is estimated from the data, the GCV continues to decrease even after 29 peaks are estimated (Fig. 5, solid lower trace, closed symbol)! This suggests that the GCV minimum estimate of the number of peaks is seriously biased.

One explanation for this bias stems from the computation of degrees of freedom used in the GCV, df=n-m-2, where m is the number of peaks detected. For each peak, two degrees of freedom are lost; one for the peak amplitude p_j, and one for the peak location within the series. Thus, the GCV ought to be penalized at a higher rate, by setting df=n-2m-2. With this approach, a more satisfactory answer for number of peaks is obtained in the simulations (Fig 5, middle trace). Now, the GCV-minimum number of peaks is 6, closer to the true value, 5, with a much more narrowly defined range of acceptable values. With the decay rate k fixed at the true value, the GCV minimum occurred at 7 peaks.

In another context, but using linear splines, Friedman (12) has suggested that the appropriate formula for df is n-3m-2, where the factor, 3, is a rough estimate of the actual number of df lost due to chosing the independent variable which reduces SS the most, rather than a random one. We may test this idea with simulations by evaluating the expected value, $E(SS_m)$ as a function of the number of peaks, m. Normalizing by the total number of

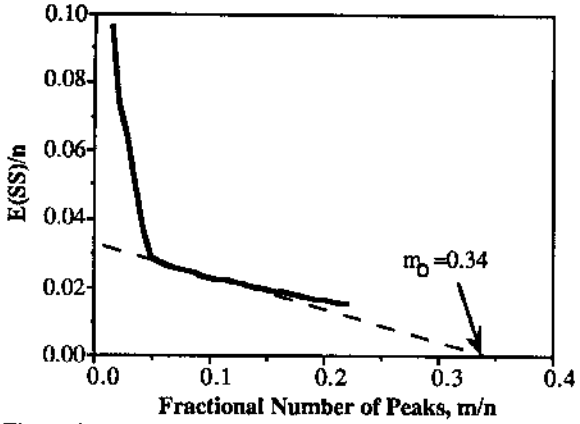

Figure 6. Expected value of residual sum of squares versus number of peaks in model, calculated from same simulations as in Figure 5. In the region of the true model, ideally, $E(SS)/n \approx \sigma^2(1-fm/n)$, so that the X intercept of the tangent line m_0, can be used to solve for f. For m_0=0.34, the penalty factor f is roughly 3.0, suggesting that the correct formula for GCV should be $SS/(n-3m-2)^2$.

measuremens, n, we may determine the correct value for the penalty factor, f from the tangent to the curve in Figure 6. In our example we see that f=3.0 is quite appropriate. We believe this result (f=3.0) should be reasonably general for many data sets of this type with about 100-200 observations and less than about 20 peaks, although the correct value for the factor f could be determined in each situation by a simulation of the type shown in Fig. 6.

DISCUSSION

We have presented a new, simplified strategy for estimating the number of secretory peaks or events from a series of hormone concentration measurements. The method accounts for first-order linear clearance of the hormone from the blood. Without making assumptions about either the intrinsic measurement error or the rate of decay, the method estimates the number of discrete events giving rise to the data. As such, this method is about as general as Veldhuis' "multiparameter deconvolution" approach (13) or the numerical deconvolution approach used by O'Sullivan (7). One detail left out of our model is the "rise-time" for each pulse, used by both (13) and (7). Some evidence of the need for this detail can be found even within the present example, wherein the rise of the peak seems to be spread over two or three intervals. An obvious way to include the "rise-time" is to change the spline basis elements, to rise gradually from 0 to 1 over the space of 1, 2 or 3 time intervals. Then, repetition of the above analysis would allow the selection of the GCV minimal rise time along with estimates of the other parameters. We do not see any conclusive evidence of the need for a second decay component, although extensive pharmacokinetic experiments have suggested its presence in LH clearance (14). It is possible that the absence of a second component has been compensated in our analysis, by the presence of LH secretion in the intervals between the 5 major peaks.

In contrast to the approaches taken by (7) and (13), our method is computationally fast, requiring only about 1 minute of computation time for analysis of the example using MATLAB on a Macintosh II. Both (7) and (13) require substantial computing time, between 12 and 20 hours on a SUN workstation, or a fairly extensive computation on an IBM mainframe. The main advantage of the short computation time is in making simulation feasible, even as a standard part of the analysis of any single data set.

A similar problem has been treated in the seismological engineering literature by Mendel, Kormylo and others (15, 16). There, investigators are interested in reconstructing the sudden changes with distance of the impedance of geologic formations. The observed data may be represented as a convolution of an impulse sequence with the characteristic response of each pulse. The use of maximum likelihood techniques with a Bernoulli-Gaussian model gives rise to a reasonable estimation procedure. However, the seismologist is primarily interested in reconstructing the underlying strata of rock, that is, the location of particular pulses in the reconstructed series. The endocrinologist is primarily interested, not in the individual peaks per se, but in peak frequency, or some overall measure of pulsatility. Thus, such engineering methods are not entirely applicable. In particular, our primary focus is on peak frequency or number of peaks, in the face of an unknown clearance rate, rather than the exact temporal location of any particular peak.

A major conclusion of this exercise is that the number of estimated peaks is highly correlated with the estimated decay rate. Second, the GCV optimal number of peaks can be heavily biased, especially when the decay rate is unknown. Third, even with the corrected GCV, there can still be substantial uncertainty in the estimated number of detected peaks. We are therefore surprised by claims of (13), that all the unknown parameters of the model: rise-time, decay rate, as well as number, location, and height of secretory peaks can be well-determined by a "multi-parameter deconvolution" method on similar data. We suspect this is only possible if the number of peaks is constrained a priori to some particular value, or range of values.

We suggest the following ideas for future work

1) Use of replicate assay measurements can, to some extent, prevent over-estimation of the number of peaks. One could match the residual mean square from the peak detection technique to the assay measurement error, and assure that the former does not become too small. Assay error is but one component of the error term in our model; biological and system variability should also be included.

2) Much work has been presented in the physiological literature, characterizing the behavior of peak detection algorithms (4). There, workers are often concerned with "false-negative" and "false-positive" rates. Frequently, ad hoc peak-detection algorithms need to be "calibrated" on white-noise to validate the nominal significance of detected peaks. One significant remaining problem in such approaches is that the false-positive rate for these methods may differ when signal is present from the rate on signal-free white noise. We feel that a more fruitful approach is to examine the bias in estimates of number of peaks, and then find minimum-bias, minimum-variance estimates. The identity of any particular peak is seldom critical.

3) Most applications of hormone pulse counting techniques might not actually require the "truth" regarding the underlying number of secretory events. Rather, as with much clinical research, comparisons between two patient groups (treated and untreated, healthy and disease) are often more relevant. Although we may not determine number of pulses precisely in the face of uncertain clearance rate, we can surely evolve a comparative index. For example, the entire curve relating clearance rate to number of peaks for a normal group (Figure 4, averaged over subjects) may be compared to a similar curve for a treated group. Relevant comparisons may be made at predetermined point, along a predetermined range of that curve, at the GCV optimal location, or using the entire length of the curve. We suspect that establishing an appropriate, statistically stable index of pulsatility, suitable for comparison of individuals and groups of patients, will be of greater importance than more elaborate deconvolution techniques.

ACKNOWLEDGEMENT

The authors are grateful to Dr. A.D. Genazzani who provided the data used in the example.

REFERENCES

1. Clifton, DK, Steiner, RA, Cycle detection: a technique for estimating the frequency and amplitude of episodic fluctuations in blood hormone and substrate concentration, Endocrinology, 112, 1057-1064, 1983
2. Merriam, GR, Wachter, KW, Algorithms for the sutdy of episodic hormone secretion ,Am. J. Physiol., 243, E310-E318, 1982
3. Oerter, KE, Guardabasso, V, Rodbard, D, Detection and characterization of peaks and estimation of instantaneous secretory rate for episodic pulsatile hormone secretion, Comput. Biomed. Res., 19, 170, 1986
4. Veldhuis, JD, Carlson, ML, Johnson, ML, The pituitary gland secretes in bursts: appraising the nature of glandular secretory impulses by simultaneous multiple-parameter deconvolution of plasma hormone concentrations. Proc Natl Acad Sci USA, 84(21), 7686-90, 1987.
5. Van Cauter, E, Estimating false-positive and false-negative errors in analyses of hormonal pulsatility, Amer. J. Physiol, 256(Endocrinol. Metab. 17), E786-E794, 1988
6. Santen, RJ, Bardin,CW, Episodic lutenizing hormone secretion in man. Pulse analysis, clinical interpretation, physiologic mechanisms, J. Clin. Invest., 52, 2617-2628, 1973
7. O'Sullivan, F, O'Sullivan, J, Deconvolution of Episodic Hormone Data: An analysis of the role of season on the onset of puberty in cows, Biometrics, 44, 339-353, 1988
8. Diggle, PJ, Zeger, SL, A Non-Gaussian Model for Time Series with Pulses, JASA, 84(406), 354-359, 1989,
9. Genazani, AD, Forti, G, Maggi, M, Milloni, M, Cianfanelli, F, Guardabasso, V, Toscano, V, Serio, M, Rodbard, D, Pulsatile Secretion of LH in Agonadal Men Before and During Testosterone Replacement Therapy, J. Andrology, submitted,1989
10. Thisted, RA, Elements of Statistical Computing, 1988, Chapman and Hall, New York
11. Craven, P, Wahba, G, Smoothing Noisy Data with Spline Functions, Numerische Mathematik, 31, 377-403, 1979
12. Friedman, JH, Silverman, BW, Flexible, Parsimonious Smoothing and Additive Modeling, Technometrics, 31(1), 3-22, 1989
13. Urban, RJ, Johnson, ML, Veldhuis, JD, In Vivo Biological Validation and Biophysical Modeling of the Sensitivity and Positive Accuracy of Endocrine Peak Detection. I. The LH Pulse Signal, Endocrinology, 124(5), 2541-2547, 1989
14. Veldhuis, JD, Fraioli, F, Rogol, AD, Dufau, ML, Metabolic clearance of biologically active luteinizing hormone in man, J. Clin. Invest., 77, 1122-1128, 1986
15. Mendel, JM, White-Noise Estimators for Seismic Data Processing in Oil Exploration, IEEE Trans. on Automatic Control, AC-22(5), 694-706, 1977
16. Kormylo, JJ, Mendel, JM, Maximum Likelihood Detection and Estimation of Bernoulli-Gaussian Processes, IEEE Trans. on Information Theory, IT-28(3), 482-488, 1982
17. Moler, C., Herskovitz, S, Little, J, Bangert, S. MATLAB for Macintosh Computers. The Math Works, Inc. 20 North Main St., Sherborn, MA 01770, 1987.

APPENDIX

The following is code for MATLAB (17) on a Macintosh II, which will implement the stepwise algorithm discussed in the text.

```
lx=length(y)
dat1=log(y);
%peaks stores a flag for inclusion in the model
peaks=zeros(lx,1);
peaks(1)=1;
newpeak=1;
%desx stores design vectors for all possible peaks
desx=tril(ones(lx,lx));
table=0;
for run=1:35
run
%Sequential knots
%are added to a linear spline.
%Knots are added according to
%the max. component of SS gradient
%x contains the columns for the knots already
%in the model
% k= 1 or 0 whether or not slope is parameter
k=1;
% build the design matrix x, including all variables for
x=desx(:,peaks);
npeaks=sum(peaks);
if k==1
  x=[x (1:lx)'];
end
%perform regression with new design matrix
beta=(x'*x)\(x'*dat1);
ypred=x*beta;
resids=dat1-ypred;
beta2=zeros(lx,1);
beta2(peaks)=beta(1:npeaks);
ss=resids'*resids;
plot ([dat1+10 ypred+10 beta2+10 resids+8 ])
%if slope is included in regression, get it
if k==1
  slope=beta(npeaks+1);
end;
table(run,1)=npeaks;
table(run,2)=slope;
table(run,3)=ss;
table(run,4)=0;
table(run,5)=newpeak;
table(run,6)=ss/(lx-npeaks-k);
table(run,7)=1000*ss/(lx-npeaks-k)/(lx-npeaks-k);
%now set up new trial knots
%dss will be vector of delta ss for each unused variable
dss=zeros(dat1);
betanew=zeros(dat1);
trialx=desx(:,1-peaks);
txresids=trialx'*resids;
dss=zeros(lx,1);
dss(1-peaks)=txresids;
%Now find largest reduction in ss
[zz,i]=sort(-dss);
peaks(i(1))=1;
newpeak=i(1);
%Now go back to include newpeak in regression
end;
```